Corpo a corpo con l'inglese della corpus linguistics, anzi, della linguistica dei corpora

Manuel Barbera, Carla Marello (Università di Torino)

Corpo a corpo con l'inglese della *corpus linguistics*, anzi, della linguistica dei *corpora*

Abstract
La terminologia della linguistica dei *corpora* è oggi dominata dall'inglese perché fa largo impiego di termini usati in linguistica computazionale, a sua volta estremamente anglicizzata, e perché i primi grandi *corpora* elettronici sono stati composti da testi in lingua inglese e i primi studi relativi sono stati condotti in Gran Bretagna o presso gli anglisti scandinavi.
In questo testo, partendo da scritti italiani di linguistica dei *corpora*, si riflette sull'utilità sia dei prestiti adattati, ad es. *tokenization/tokenizzazione*, sia dei prestiti non adattati, e si sostiene l'opportunità di non avallare traduzioni imprecise.

1. C'è *corpus* e *corpus*

La terminologia della linguistica dei *corpora*[1], a dispetto del latinismo *corpus/corpora*[2], è largamente dominata oggi dall'inglese per almeno due motivi:

a) fa largo impiego di termini usati in linguistica computazionale, a sua volta estremamente anglicizzata;

b) i primi *corpora* e i più grandi, non legati alla produzione di un singolo autore, sono (stati) i *corpora* di lingua inglese e in Gran Bretagna o presso gli anglisti scandinavi si sono sviluppati gli studi più numerosi e vivaci di *corpus linguistics*.

Quando si traccia la storia di questa branca della linguistica, c'è la diffusa tendenza a parlare di studi di *corpus linguistics* per tutti gli studi basati su dati empirici anche prima di quaranta anni fa[3]. Svartvik 1992, p. 7, risale addirittura a Otto Jespersen, ma come osserva Rossini Favretti 1998, p. 47 : «Ciò che caratterizza la costituzione dei *corpora* in *corpus linguistics* è che questi non sono limitati alle sequenze o agli enunciati 'notati' dal linguista. I *corpora* più recenti sono 'monitorati' per recepire i dati linguistici secondo parametri che tendono a ridurre al minimo le possibilità di scelta del soggetto e mirano alla massima estensione ed oggettività. Se l'oggettività non è raggiungibile, l'estensione diviene il punto di riferimento ai fini della scientificità dell'indagine».

[1] Questo articolo è stato scritto in collaborazione dai due autori e di entrambi è il paragrafo conclusivo; è di Manuel Barbera il paragrafo 3 e sono di Carla Marello i paragrafi 1 e 2. Le ricerche durante le quali sono maturate queste riflessioni sulla terminologia si inseriscono nel progetto Cofin 2001 *Italiano antico e informatica*, coordinatore nazionale Lorenzo Renzi e FIRB 2001 *L'italiano nella varietà dei testi. L'incidenza della variazione diacronica, testuale e diafasica nell'annotazione e interrogazione di corpora generali e settoriali*, coordinatore Carla Marello.

[2] In italiano *corpus* con il senso di raccolta completa e ordinata di opere letterarie o giuridiche viene datato dai dizionari come attestato tra la seconda metà del XIX e gli inizi del XX; come campione rappresentativo di una lingua studiato dal linguista (e quindi non necessariamente formato di testi interi) si afferma più tardi attraverso l'innesto dell'uso inglese sulla tradizione filologica locale.Nel 1993 il *New Shorter Oxford English Dictionary* registra come della metà del ventesimo secolo il significato specialistico «A body of spoken or written material on which a linguistic analysis is based». Sabatini-Coletti 2003 alla voce «**corpus:** 2. ling. Raccolta di brani, singoli enunciati o altri dati linguistici, che vengono analizzati per definire la struttura di un sistema linguistico • campione» dà come etimologia « • voce lat., "corpo"; nell'accez. 2. entrata dall'ingl. a. 1968».

[3] Si veda tuttavia l'osservazione di Leech 1991, pp. 8-9 che ritiene più opportuno partire dalla "rifondazione" degli anni Sessanta ad opera di Quirk, Francis e Kučera.

Il termine *corpus linguistics* tuttavia si è affermato nell'ambito scientifico di lingua inglese solo una ventina di anni fa[4] e più tardi in Italia.

Ad esempio, nel 1991 quando già la *corpus linguistics* ha una sua visibilità internazionale e nazionale[5], Diego Marconi in un contributo rivolto a insegnanti italiani annovera la «linguistica descrittiva basata su *corpora*» fra le applicazioni delle tecniche «non intelligenti» di elaborazione del linguaggio naturale e non ricorre ancora al termine.[6] Cinque anni dopo i tempi sono maturi per usare la locuzione nel titolo di un capitoletto di un libro divulgativo, osservando che c'è sempre stata una linguistica basata sullo spoglio di materiali linguistici, ma che la linguistica dei *corpora* è una linguistica dei *corpora* elettronici.[7]

Con il già citato libro di Rossini Favretti 1998, la linguistica dei *corpora* approda in un manuale universitario di linguistica applicata, ma stenta a penetrare nei dizionari di linguistica, perché si porta dietro un problema teorico. Come sempre succede coi termini, se sono controversi è perché ciò che indicano è controverso.

Per molti la linguistica dei *corpora* continua ad essere un modo di fare linguistica e di provarla su *corpora*: per questi *corpus linguistics* è un modo scorciato di denominare la cosiddetta *corpus-based linguistics*, non serve quindi a individuare una disciplina autonoma. Chi invece sostiene una *corpus-driven linguistics*, una linguistica a partire dal *corpus*,[8] è più incline a considerarla una branca a sé e vuol sottolineare come per conoscere davvero la lingua in uso sia necessario creare *corpora* molto estesi e bilanciati, o mettere in relazione una serie di *corpora* nati separatamente ed

[4]Una delle prime attestazioni in un titolo è Aarts and Mejis eds. 1984.

[5] McEnery e Wilson 1996 osservano che fino al 1965 gli studi ascrivibili a quella che oggi si chiama *corpus linguistics* erano 10 e che in seguito ogni cinque anni sono raddoppiati: nel periodo 1986-1991 sarebbero stati ben 320.

[6] Marconi 1991, p. 102

[7] «Nonostante *corpus linguistics* ci arrivi dall'inglese, è però una bella combinazione che *corpus-corpora* sia latino, non tanto per la tradizione filologica a cui risale il termine, quanto per la curiosa coincidenza insita nel fatto che i primi passi nella pratica degli spogli elettronici sono stati fatti proprio in Italia su un insieme di testi in latino (cfr. Busa 1951), l'opera di San Tommaso d'Aquino » Marello 1996, pp. 167-172.

[8] Un composto con *–driven* che i linguisti italiani hanno dovuto affrontare nello stesso periodo è stato quello in Head-driven Phrase Structure Grammar di Pollard e Sag nel 1994. Non l'hanno di fatto tradotto e negli ambienti linguistici si è usato HPSG. Allegranza e Mazzini, nel primo manuale universitario italiano dedicato alle grammatiche ad unificazione (2000, p. 185), hanno scelto una traduzione terminologicamente molto appropriata, «grammatica a struttura sintagmatica proiettata dalla testa», ma per comodità e consuetudine il più delle volte hanno usato la sigla inglese. Nel caso di *corpus-driven* non si può parlare di 'proiezione', termine tecnico legato alla teoria X-barra, proposta da Chomsky nel 1970.

elaborare finestre di interrogazione adeguate e come tutto ciò richieda conoscenze linguistiche e informatiche specifiche.

Quando Sinclair 2000, p.29, afferma «per i primi trenta anni in cui ho lavorato nella CL (= *corpus linguistics*) quasi nessuno se ne è accorto», riporta in modo indiretto questa situazione. Del resto lui stesso, uno dei più decisi paladini della linguistica a partire dal *corpus*, curando un libro del 1987, *Looking Up. An account of the COBUILD Project in lexical computing*, oggi considerato uno dei testi fondanti della linguistica dei *corpora*, usa *corpus* in combinazione con molte parole (traggo dall'indice analitico *composition, correction, data, development, processing*), ma non premesso a *linguistica*. Può essere un caso, ma riteniamo piuttosto che sia quanto accadeva ancora sul finire degli anni Ottanta. Chi si occupava di linguistica dei *corpora* si considerava un linguista computazionale e se mai polemizzava con chi riduttivamente lo tacciava di fare linguistica quantitativa. Aveva molto da polemizzare, perché l'attacco poteva venire sia dai generativisti, sia dai cultori dei modi "intelligenti" di elaborare il linguaggio naturale. In un certo senso allora si spendeva più tempo a stabilire che cosa doveva essere un *corpus*, in che cosa differiva da una base di dati lessicali, che non a dare un nome alla ricerca che si faceva.

Nella culla allora più nota della *corpus linguistics* italiana, la Pisa dell'Istituto CNR di Linguistica Computazionale creato nel 1978, la Pisa della Scuola Normale, si usava, scrivendo in italiano, *corpus* con estrema parsimonia, forse perché il pubblico colto associava la parola *corpus* a un insieme di testi non su supporto elettronico. Qualche esempio: nel 1979 Paola Barocchi, trattando specificamente di quella che noi oggi diremmo *corpus linguistics* al servizio della lessicografia tecnica, accennava all' «applicazione della memorizzazione elettronica ai dati e documenti storico-artistici» e Giovanni Nencioni nel 1984, parlando di lessici ad una conferenza su "Automatic Processing of Art History Data and Documents", usa parole come 'archivi testuali', 'processo di automazione' di tali archivi. E solo una volta accenna a un sistema per «interrogare archivi di testi al livello di forma delle parole, rintracciando nel *corpus* una parola data, calcolando la sua

frequenza, fornendo contesti di determinata lunghezza, cercando occorrenze di parole nello stesso contesto» (1984, p. 23).

Per qualche tempo parallelamente a *banca di dati* si affermò anche *text bank, banca testuale*[9]

Secondo Spina 2001, p. 64, da circa tre decenni *corpus* ha assunto l'accezione di insieme di dati linguistici pronti per essere elaborati da un computer, anzi di «raccolta strutturata di testi in formato elettronico che si assumono rappresentativi di una data lingua o di un suo sottoinsieme, mirata ad analisi di tipo linguistico». L'unione di questa accezione di *corpus* a *linguistics* è però, come abbiamo detto, più recente, ma sarebbe, secondo alcuni, già superata dalla *linguistica delle risorse.*[10]

I dizionari monolingui generali di italiano continuano a trascurare il fatto che i *corpora* in linguistica oggi sono elettronici: si veda lo Zingarelli 2004 (2003) «(ling.) Campione rappresentativo di una lingua che il linguista prende in esame» e il Sabatini-Coletti 2003 nell'etimologia specifica come l'uso linguistico viene dall'inglese più che dal latino, ma non accenna al supporto elettronico.[11]

Se si lancia la ricerca della parola *corpora* in tutto il testo del *Grande Dizionario Italiano dell'Uso GDU* di Tullio De Mauro (UTET, versione elettronica aggiornata al 2003), il solo risultato[12] che si ottiene è una citazione dantesca all'accezione ottava di **Mestiere**: «ufficio funebre, esequie, funerale: *tutti li dolorosi mestieri che a le corpora de li morti s'usano di fare* (Dante)».

[9] Troviamo il termine usato ad esempio da Knowles 1990, p.1655 in un articolo enciclopedico in cui oggi si userebbe di sicuro *corpus/corpora.*

[10] «Secondo il vecchio paradigma lo studio di unità diverse di un *corpus* necessitava la messa a punto di processi specifici diversi per ogni fenomeno studiato. (…) Una risorsa linguistica ammette l'annotazione delle informazioni relative a diversi livelli di analisi (…) tutti espressi in un linguaggio di *mark-up.*» Ferrari 2000, pp. 24-25.

[11] Vedi più sopra nota 2.
Si noti per altro che pure i dizionari inglesi sottolineano questo aspetto della natura computerizzata del *corpus* nella moderna linguistica abbastanza tardi. Nel 1998 il *New Oxford Dictionary of English* parla esplicitamente nella prefazione di *corpus analysis* e di *evidences* trovate «using computational tools to analyse the data in the British National Corpus» e ha poi nella definizione di *corpus* il *subsense* : « a collection of written or spoken material in machine-readable form, assembled for the purpose of studying linguistic structures, frequencies, etc. ».

[12] Infatti ormai De Mauro dà *corpus* come sostantivo maschile invariabile e non segnala il plurale *corpora.*

2. *Plain* or *raw corpus*

Il sintagma preposizionale 'del *corpus*' o 'dei *corpora*' seguendo il sostantivo a cui si riferisce rende piuttosto difficili i tentativi di rendere in italiano le specificazioni e modificazioni di *corpus e* di *corpus linguistics*. Quanto al nome di chi intraprende questi studi, *linguista dei corpora* o dei *corpus* (se si volesse ritenere la parola invariabile, come suggerisce il dizionario di De Mauro) non si è affermato (Spina 2001, ad esempio, usa l'inglese *corpus linguist*). Infatti volendo aggettivare *il linguista dei corpora* o si usano aggettivi che possono stare in posizione prenominale o altrimenti si deve rinunciare: *? il linguista italiano dei corpora, ? il linguista dei corpora italiano*.

Anche la modificazione o specificazione di *corpus* si presenta problematica: non ha derivati in italiano, a fronte della possibilità dell'inglese di trasformarlo in aggettivo; il suo plurale *corpora* è difficile e già ora gli studenti seguono, inconsapevolmente, il suggerimento di De Mauro di considerare *corpus* invariabile. *Monitor corpus, reference corpus, training corpus* sono stati tradotti rispettivamente 'corpus di monitoraggio', 'corpus di riferimento' e 'corpus di allenamento'. *Text corpora* sono i *corpora* di lingua scritta,[13] *spoken corpora,* 'la parte parlata di un corpus altrimenti composto di testi scritti' viene resa in italiano come '*corpora* orali o di parlato'.

Plain o *raw corpus* non è di solito tradotto letteralmente anche se i catalani l'hanno fatto e parlano di *corpus en brut*. In italiano in questo caso più che avventurarsi in '*corpus* grezzo o piano' si ricorre alla negazione '*corpus* non annotato', che contrappone questo tipo di *corpus* a quelli annotati per parte del discorso o altro.

[13] Anche se in ambito tedesco e italiano, dove molti linguisti sono arrivati alla linguistica dei *corpora* dalla linguistica testuale oltre che dalla lessicografia, *text corpora* è sempre stato inteso come '*corpora* di testi scritti' e non semplicemente come '*corpora* di materiali scritti'. Cfr. Bergenholtz e Schaeder 1979.

3. *Concordanziale* e *tokenizzare*

La fortuna di un termine tecnico dipende anche, in lingue ricche di morfologia derivazionale, quale è l'italiano, dalla sua capacità di dar origine ad aggettivi denominali, a verbi denominali, a nomi deverbali.

Le concordanze[14] sono sempre state uno dei prodotti più conosciuti della linguistica dei *corpora.* Insieme a *corpus* il termine è un'altra eredità della filologia: è assai poco trasparente per gli studenti che tendono a confonderlo con l'*accordo,* detto anche in qualche grammatica *concordanza*. In compenso intorno alla parola troviamo *concordanziale* e *concordatore,*[15] inteso come 'colui che fa le concordanze' e non come 'programma informatico per fare concordanze' come invece significa *concordancer* in inglese.

Passaggio indispensabile per la linguistica dei *corpora* è l'individuazione dei *token.* Ecco come Peirce definisce *type* e *token* nel 1906, con un'esemplificazione che sembra fatta apposta per la linguistica dei *corpora*:

> «A common mode of estimating the amount of matter in a MS. or printed book is to count the number of words. There will ordinarily be about twenty *the's* on a page, and of course they count as twenty words. In another sense of the word 'word', however, there is but one word 'the' in the English language; and it is impossible that this word should lie visibly on a page or be heard in any voice, for the reason that it is not a Single thing or Single event. It does not exist; it only determines things that do exist. Such a definitely significant Form, I propose to term a *Type*. A Single event which happens once and whose identity is limited to that one happening or a Single object or thing which is in some single place at any one instant of time, such event or thing being significant only as occurring just when and where it does, such as this or that word on a single line of a single page of a single copy of a book, I will venture to call a *Token* (…) In order that a Type may be used, it has to be embodied in a Token which shall be a sign of the Type, and thereby of the object the Type signifies. I propose to call such a Token of a Type an *Instance* of the Type.» ('Prolegomena to an Apology for Pragmaticism', CP 4.537, 1906)

I *types* non sono i lemmi, ma i nomi delle classi di tutti i *tokens* che hanno la stessa forma; è perciò fonte di confusione non distinguere anche terminologicamente la *tokenizzazione* di un testo dalla sua *lemmatizzazione*. Si potrebbe usare *segmentazione,* ma sarebbe generico rispetto alla ricchezza informativa di *tokenizzazione*. Si potrebbe adottare *occorrenza* al posto di *token,* ma la parola non si

[14] « Le concordanze sono una lista delle occorrenze di una o più forme, ciascuna mostrata all'interno del contesto in cui compare nel corpus » Spina 2001, p. 127. Il termine originariamente indicava «libro che dà [...] tutti i passi che si riferiscono ad un determinato pensiero od oggetto (concordanza reale). Specialmente notevoli le "concordanze" della Bibbia» Panzini 1905, *sub vocem.*

[15] Troviamo queste parole usate soprattutto da Savoca . Cfr. Savoca 2000

presta a creare utili verbi denominali da cui ricavare i deverbali indispensabili per nominare i processi di elaborazione dei testi.

A favore dell'adozione di *tokenizzazione,* vi è poi il mantenere lo stesso termine per processi che interessano campi disciplinari distinti come la semiotica, la filosofia del linguaggio e appunto la linguistica. Campi che la definizione di Peirce, inventore del significato specifico di *token* qui discusso, tiene uniti ricorrendo per la propria argomentazione filosofica ad un esempio di àmbito linguistico. Ci voleva la "stupidità" del computer perché i linguisti si accorgessero della sensatezza della proposta di Peirce.

Il termine italianizzato nella parte derivativa, ma lasciato con la *k* dell'inglese *tokenize/tokenization*, è facile da pronunciare e scrivere e viene riconosciuto facilmente anche da un non italiano, in quanto è un internazionalismo della linguistica dei *corpora*. Spina (2001, p. 108) usa *tokenizzazione,* precisando:

> « L'intuitività della nozione di *token* non deve trarre in inganno (…) la regola abitualmente associata ai software di calcolo delle frequenze prevede che parole siano le porzioni di testo precedute e seguite da uno spazio e/o da un apostrofo (…) L'italiano presenta però una situazione molto più complessa riguardo alla segmentazione di un testo (…) come nel caso dei clitici e delle preposizioni articolate» Spina (2001, p. 108)

Savoca tokenizza nei suoi vari volumi di concordanze, ma come molti altri considera l'operazione un preliminare senza nome della lemmatizzazione.

Dati i limiti di spazio della presente comunicazione, crediamo che la miglior cosa sia dare una definizione ostensiva dei vari passaggi che un testo deve fare per arrivare a far parte di un *corpus,* utilizzando un brano del *Tesoretto* di Brunetto Latini (vv. 113-134), originariamente èdito nei *Poeti del Duecento* di Contini (tav. 1a), spogliato anche per l'OVI (tav.1b) e confluito infine nel *Corpus Taurinense* [16] (tav. 1c).

[16]Il *Corpus Taurinense* è un sottoinsieme fiorentino duecentesco della raccolta di testi del *TLIO* (*Tesoro della lingua italiana delle origini*), tokenizzato ed annotato morfosintatticamente (ossia *POS-tagged*, 'annotato per parti del discorso'). È interrogabile in ambiente UNIX (Solaris o Linux) con il CWB (Corpus Work Bench) dell'IMS di Stoccarda (cfr. Christ / Schulze 1996), ed è attualmente consultabile in rete nel sito dell'IMS. Nel CT sono possibili

Nella Tavola 1 si vede la differenza tra un testo editorialmente "normale" (ma già più raffinato della media: si veda la distinzione nell'uso dell'apostrofo su gruppi tonici ed atoni) e due testi con diversi gradi di tokenizzazione, funzionali a due diversi progetti di elaborazione elettronica, l'uno lessicografico e l'altro di linguistica dei *corpora*.

testo non tokenizzato Contini, *Poeti del Duecento*	**tokenizzazione parziale** testo OVI	**tokenizzazione totale** *Corpus Taurinense* non annotato
Lo Tesoro conenza. Al tempo che Fiorenza froria, e fece frutto, sì ch'ell'era del tutto la donna di Toscana (ancora che lontana ne fosse l'una parte, rimossa in altra parte, quella d'i ghibellini, per guerra d'i vicini), esso Comune saggio mi fece suo messaggio all'alto re di Spagna, ch'or è re de la Magna e la corona atende, se Dio no·llil contende: ché già sotto la luna non si truova persona che, per gentil legnaggio né per altro barnaggio, tanto degno ne fosse com' esto re Nanfosse.	Lo Tesoro conenza. Al tempo che Fiorenza froria, e fece frutto, sì ch' ell' era del tutto la donna di Toscana (ancora che lontana ne fosse l' una parte, rimossa in altra parte, quella d' i ghibellini, per guerra d' i vicini), esso Comune saggio mi fece suo messaggio all' alto re di Spagna, ch' or è re de la Magna e la corona atende, se Dio no· llil contende: ché già sotto la luna non si truova persona che, per gentil legnaggio né per altro barnaggio, tanto degno ne fosse com' esto re Nanfosse.	Lo Tesoro conenza . A ÷l tempo che Fiorenza froria , e fece frutto , sì ch' ell' era de ÷l tutto la donna di Toscana (ancora che lontana ne fosse l' una parte , rimossa in altra parte , quella d' i ghibellini , per guerra d' i vicini) , esso Comune saggio mi fece suo messaggio a ÷ll' alto re di Spagna , ch' or è re de la Magna e la corona atende , se Dio no· lli ÷l contende : ché già sotto la luna non si truova persona che , per gentil legnaggio né per altro barnaggio , tanto degno ne fosse com' esto re Nanfosse .
Apici senza spazio su gruppi proclitici, con spazio su gruppi tonici; scempiamento su assimilazione in clisia con punto in alto attaccato; punteggiatura attaccata.	*Apici sempre separati, punteggiatura in genere separata, separati i punti di clisia.*	*Tutti gli apici separati, tutti gli interpuntemi separati, tutti i punti di clisia separati, tutti i clitici grafici separati.*

Tavola 1 a b c

tokenizzazione + mark-up *Corpus Taurinense* non annotato ma con mark-up testuale	
@BrunettoLatini@@Tesoretto@@@Did %001 0175 &V [...] Lo Tesoro conenza . A ÷l tempo che Fiorenza froria , e fece frutto , sì ch' ell' era de ÷l tutto la donna di Toscana (ancora che lontana ne fosse l' una parte , 0180 rimossa in altra parte , quella d' i ghibellini ,	per guerra d' i vicini) , esso Comune saggio mi fece suo messaggio a ÷ll' alto re di Spagna , ch' or è re de la Magna e la corona atende , se Dio no· lli ÷l contende : ché già sotto la luna non si truova persona che , per gentil legnaggio né per altro barnaggio , tanto degno ne fosse com' esto re Nanfosse .
markup: @autore @@titolo @@@genere %capitolo $pagina &v verso	

Tavola 2

---

ricerche per diverse categorie di fenomeni vuoi linguistici (forma, lemma, POS, categorie morfologiche), vuoi filologici (correzioni editoriali, genere letterario, prosa/verso) in qualsivoglia combinazione seriale.

Nella Tavola 2 al medesimo testo è aggiunto *mark-up* (ovvero 'marcatura ipertestuale') per introdurre informazioni "non lineari" in formato lineare (ossia riportate nel medesimo formato di "stringa di caratteri" del "testo" cui sono sovrapposte) e quindi leggibile dal computer.

tokenizzazione + markup + POS-tagging
Corpus Taurinense con mark-up testuale e POS-tagging

```
Lo_lem=lo,60,0,4,6,0,0 Tesoro_lem=tesoro,20,0,4,6,0,0 conenza_lem=cominciare,111,3,0,6,0,0
    ._lem=stop,70,0,0,0,0,0
A_lem=a,56,0,0,0,0,0 ÷l_lem=il,60,0,4,6,0,0 tempo_lem=tempo,20,0,4,6,0,0
    che_(lem=che,36,0,4;5,6;7,0,0);(lem=che,51,0,0,0,0,0);(lem=che,35,0,4;5,6,0,0);(lem=che,40,0,4;5
    ,6,0,0);(lem=che,32,0,4,6,0,0) Fiorenza_lem=firenze,21,0,5,6,0,0
froria_lem=fiorire,112,3,0,6,0,0 ,_lem=comma,71,0,0,0,0,0 e_lem=e,50,0,0,0,0,0 fece_lem=fare/-
    si/,113,3,0,6,0,0 frutto_lem=frutto,20,0,4,6,0,0 ,_lem=comma,71,0,0,0,0,0
sì_lem=sì,45,0,0,0,8,0
    ch'_(lem=che,36,0,4;5,6;7,0,0);(lem=che,35,0,4;5,6,0,0);(lem=che,51,0,0,0,0,0);(lem=che,40,0,4;5
    ,6,0,0);(lem=che,32,0,4,6,0,0) ell'_lem=ella,37,3,5,6,0,0 era_lem=essere,212,3,0,6,0,0
    de_lem=di,56,0,0,0,0,0 ÷l_lem=il,60,0,4,6,0,0
    tutto_(lem=tutto,32,0,4,6,0,0);(lem=tutto,45,0,0,0,8,0);(lem=tutto,20,0,4,6,0,0);(lem=tutto,26,0
    ,4,6,8,0);(lem=tutto,51,0,0,0,0,0)
la_lem=la,60,0,5,6,0,0 donna_lem=donna,20,0,5,6,0,0 di_lem=di,56,0,0,0,0,0
    Toscana_lem=toscana,21,0,5,6,0,0
```

tokenizzazione + markup + POS-tagging + disambiguazione
Corpus Taurinense con mark-up testuale e POS-tagging disambiguato (versione finale)

```
Lo_lem=lo,60,0,4,6,0,0 Tesoro_lem=tesoro,20,0,4,6,0,0 conenza_lem=cominciare,111,3,0,6,0,0
    ._lem=stop,70,0,0,0,0,0
A_lem=a,56,0,0,0,0,0 ÷l_lem=il,60,0,4,6,0,0 tempo_lem=tempo,20,0,4,6,0,0
    che_lem=che,36,0,4;5,6;7,0,0 Fiorenza_lem=firenze,21,0,5,6,0,0
froria_lem=fiorire,112,3,0,6,0,0 ,_lem=comma,71,0,0,0,0,0 e_lem=e,50,0,0,0,0,0 fece_lem=fare/-
    si/,113,3,0,6,0,0 frutto_lem=frutto,20,0,4,6,0,0 ,_lem=comma,71,0,0,0,0,0
sì_lem=sì,45,0,0,0,8,0 ch'_lem=che,51,0,0,0,0,0 ell'_lem=ella,37,3,5,6,0,0
    era_lem=essere,212,3,0,6,0,0 de_lem=di,56,0,0,0,0,0 ÷l_lem=il,60,0,4,6,0,0
    tutto_lem=tutto,32,0,4,6,0,0
la_lem=la,60,0,5,6,0,0 donna_lem=donna,20,0,5,6,0,0 di_lem=di,56,0,0,0,0,0
    Toscana_lem=toscana,21,0,5,6,0,0
```

Tavola 3

Nella Tavola 3 i primi cinque versi del testo tokenizzato e corredato di *mark-up*[17] sono stati ulteriormente "annotati morfosintatticamente" (ciò che in inglese si dice più precisamente *POS-*

[17] Il Sabatini-Coletti 2003 e lo Zingarelli 2004 riportano **mark-up** con il solo significato economico « Margine di profitto da aggiungere al costo totale di produzione per definire il prezzo di vendita di un prodotto » (Zingarelli 2004, da cui è stratta la definizione, lo data 1993; il Sabatini- Coletti a. 1987). Il GDU 2003 ammette la grafia sia con trattino sia *markup*, pl. *markups*, [der. di (*to*) *mark up* "valorizzare"] e come secondo significato dà quello informatico «[di] linguaggio informatico, che permette di segnalare attraverso marcatori o tag, le caratteristiche di un documento o di parti di esso». Nessuno dei tre dizionari suggerisce un termine italiano, che potrebbe essere eventualmente quello da noi proposto come possibile traducente, *marcatura*, in sé troppo generico, e quindi necessariamente seguito dalle necessarie specificazioni, come *ipertestuale* (che comunque è anch'essa un anglicismo...). Fra gli addetti ai lavori si sta tuttavia diffondendo come verbo corrispondente *marcappare*, con adattamento grafico in base alla pronuncia italianizzata, o spesso anche senza, *markuppare*. I testi delle tavole 2 e 3 sarebbero perciò gergalmente definiti come "marcappati". Procedimento analogo è, tra l'altro, già avvenuto con *taggare* da *tag*,entrambi registrati dal GDU. Il mantenimento (più o meno adattato) della forma inglese potrebbe essere pertanto suggerito, oltre che dai vantaggi internazionalistici e dalla maggiore stringatezza e precisione, anche dalla facilità con cui in italiano se ne possono formare derivativi.

tagging, POS= Part Of Speech). In 3a sono mantenute tutte le transcategorizzazioni che in 3b sono disambiguate.

Transcategorizzazione (e parallelamente in inglese *transcategorization*) è il termine tecnico, adottato nelle raccomandazioni europee EAGLES,[18] con cui si indica la possibile appartenenza di un *token* a più *types* e quindi, nello specifico, il fatto che gli siano attribuiti più *POS-tags* alternativi. *"Disambiguare una transcategorizzazione"* significa conservare al *token* la sola annotazione morfosintattica (*POS-tag*) richiestagli dal contesto[19].

Quanto al termine *tag*, si può usare al suo posto *annotazione* solo in qualche caso. La tendenza ad usare sempre *etichetta* sulla scorta del francese è diffusa, ma non condivisa da tutti.[20]

4. Male minore o segno di vitalità?

Nelle pagine precedenti abbiamo cercato di tracciare le origini terminologiche di una branca della linguistica che pone problemi terminologici già a partire dal suo stesso nome. I lettori si saranno accorti dei tentativi che abbiamo fatto per evitare un'accettazione "incontrollata e irriflessa"[21] dei termini stranieri: il prezzo che abbiamo dovuto pagare è stato però alto, nel senso che spesso abbiamo dovuto o rinunciare al termine italiano o mettere, per essere sicuri di farci capire almeno all'interno della cerchia degli addetti ai lavori, il termine inglese tra parentesi subito appresso al corrispondente italiano adottato.

[18] Cfr. Monachini 1996 e Monachini - Calzolari 1996.
[19] Si noti come questo uso di *disambiguare* differisca da quello che se ne fa in relazione a frasi ambigue in grammatica generativa.
[20] Si veda ad esempio Barbera 2003, nota 14: «L'inglese (cfr. ad es. Leech 1997a: 25) rende possibile distinguere tra *tag* 'categoria morfologica associata ad una determinata parola' (ad esempio 'preposizione'), *label* 'il nome o la codifica con cui un *tag* è indicato' (ad esempio "prep" o "IN") ed *adnotation* 'l'operazione od il risultato dell'applicazione dei *tags*' (ad esempio *con_prep l'_art ombrello_n*), laddove l'italiano dispone solo *di annotazione* ed *etichetta*. Io nel prosieguo cercherò di usare *etichetta* nel solo significato di 'label', ricorrendo a *tag* al posto di *annotazione* solo quando l'uso di *annotazione* nel senso di 'tag' riuscisse incongruo all'uso italiano o controindicato nel singolo contesto».
[21] Giustamente bollata in Sabatini e Scarascia Mugnozza 2003, p. 2

In generale l'adattamento fonico della parola inglese e la derivazione morfologicamente adattata (*tokenizzare*, *taggare*, ecc.) appare la soluzione forse meno elegante ma più praticata e praticabile nell'ambito della linguistica dei *corpora*. Non riuscendo a convincere gli addetti ai lavori, specie i formatori e i divulgatori, ad adottare una traduzione italiana unica e accettata (e non ci pare esistere per ora una tendenza in tal senso), l'adattamento resta la via meno dispendiosa in termini di memoria per l'utente e quindi quella con le maggiori probabilità di successo e diffusione.
L'adattamento ha i vantaggi di tutte le internazionalizzazioni, cioè viene capito anche dagli esperti non italiani[22] e implica comunque un inglobamento maggiore dei termini nelle caratteristiche della fonetica e della grafia italiane. È ben vero che viene praticato a livello fonico ma non grafico nei termini base (*markup, token*) e che solo talvolta investe pure a livello grafico i derivati più ostici (*marcappare* vs. *tokenizzare*). Di solito hanno maggior successo nel tempo gli adattamenti che non implicano un cambiamento grafico tra termine base e derivati e, se la parola base è troppo lontana dalla grafia italiana per ammettere derivazioni indolori, si ricorre piuttosto alla locuzione verbale *fare un markup, dividere in chunk.*
Il terminologo che voglia proporre termini italiani sostitutivi di quelli stranieri avrebbe più successo se si ispirasse anche alle procedure economiche (massimo rendimento col minimo sforzo) di creazione neologica praticate attraverso l'adattamento dagli italiani attivi nei vari campi specialistici; in particolare farebbe bene a considerare che nel campo dell'elaborazione informatica del linguaggio naturale operano esperti con conoscenza delle altre lingue superiore alla media e soprattutto non indifferenti alla forma delle parole in relazione al significato e frequenza d'uso.

[22] Purché l'adattamento non sia troppo radicale come in *marcappare*, poco riconoscibile per uno straniero rispetto a *markuppare*, considerato però un adattamento oneroso da alcuni italiani, che a quel punto piuttosto dicono *fare un markup*.

Bibliografia

Jan Aarts and Willem Mejis eds., *Corpus Linguistics; Recent Developments in the Use of Computer Corpora in English Language Research,* Amsterdam, Rodopi, 1984

Valerio Allegranza, Giampaolo Mazzini, *Linguistica generativa e grammatiche a unificazione,* Torino, Paravia Scriptorium 2000

Manuel Barbera, *Italiano antico e linguistica dei* corpora*: un* tagset *per ITALANT* , in Elisabeth Burr, (ed.): *Tradizione & Innovazione. Linguistica e filologia italiana alle soglie di un nuovo millennio.* Atti del VI Convegno Internazionale della SILFI, 28 Giugno - 2 Luglio 2000, Gerhard-Mercator-Universität Duisburg, Germania. Firenze: Cesati 2003

Paola Barocchi, *Introduzione al Convegno,* in *Atti del Convegno nazionale sui Lessici tecnici delle arti e dei mestieri,* Cortona 28-30 maggio 1979 Firenze, Eurografica spa, 1979, pp. 15-19

Henning Bergenholtz , Burkhard. Schaeder (Hrsg), *Empirische Textwissenschaft: Aufbau und Auswertung von Text-Corpora*, Königstein, Scriptor Verlag, 1979

Adriana Teresa Damascelli, Aurelia Martelli, *Corpus linguistics and computational linguistics: an overview with special reference to English,* Torino, CELID, 2002

Tullio De Mauro, *Grande Dizionario Italiano dell'Uso GDU* Torino, UTET, 2003 (seconda edizione elettronica aggiornata)

Giacomo Ferrari, *Livelli di analisi del testo. Due approcci a confronto,* in *Linguistica e informatica. Corpora, multimedialità e percorsi di apprendimento,* a cura di Rema Rossini Favretti, Roma, Bulzoni, 2000, pp. 15-27

Roger Garside, Geoffrey Leech, Geoffrey Sampson (a cura di), *The computational analysis of English. A corpus-based approach*, London, Longman. 1987

Roger Garside, Geoffrey Leech, Anthony McEnery, (edd.), *Corpus Annotation. Linguistic Information from Computer Text Corpora,* London - New York: Longman. 1997,

Frank Knowles, *The Computer in Lexicography*, in Franz.Joseph Hausmann., Oskar Reichmann, Herbert Ernst Wiegand, Ladislav Zgusta. (Hrsg.), *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie*, 3 voll. Berlin, New York, de Gruyter, 1989-1991, vol.II, 1990, pp. 1645-1672

Geoffrey Leech, *The state of the art in corpus linguistics,* in *English Corpus Linguistics. Studies in Honour of Jan Svartvik,* a cura di Karin Aijmer e Bengt Altenberg, London, Longman, 1991, pp. 8-29

Geoffrey Leech, *Grammatical Tagging*, in Roger Garside, Geoffrey Leech, Anthony McEnery (edd.), 1997, pp. 19-33.

Diego Marconi, *L'elaborazione del linguaggio naturale nell'ambito dell'Intelligenza Artificiale,* in *Riflettere sulla lingua,* a cura di Carla Marello e Giacomo Mondelli, Scandicci-Firenze, La Nuova Italia, pp. 101-121

Carla Marello, *Le parole dell'italiano. Lessico e dizionari,* Bologna, Zanichelli, 1996

Tony McEnery, Andrew Wilson, *Corpus linguistics,* Edinburgh, Edinburgh U.P. 1996

Monica Monachini, *Specifications for Italian Morphosyntax - Lexicon Specifications and Classification Guidelines*, Pisa, EAGLES Document (EAG-CLWG-ELM-IT/F) 1996

Monica Monachini, e Nicoletta Calzolari, *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Application to European Languages*, Pisa, EAGLES Document (EAG-CLWG-MORPHSYN/R), 1996

Giovanni Nencioni, *Intervento sui Lessici,* in *Proceedings of the Second International Conference on Automatic Processing of Art History Data and Documents*, a cura di Laura Corti e Marilyn Schmitt, Firenze, Regione Toscana, 1984, pp. 9-29

Nelleke Oostdijk, *Corpus Linguistics and the Automatic Analysis of English,* Amsterdam, Rodopi 1991

Alfredo Panzini, *Dizionario moderno delle parole che non si trovano nei dizionari comuni; con un proemio di A. Schiaffini e una appendice di ottomila voci nuovamente compilata da B. Migliorini*, Milano, Ulrico Hoepli, nona edizione, 1950; prima edizione: *Dizionario moderno: supplemento ai dizionari italiani [...]: storia, etimologia e filosofia delle parole*, Milano, Hoepli, 1905.

Charles Sanders Peirce, *Prolegomena to an Apology for Pragmaticism*, 4.537, 1906; in *The collected papers of Charles Sanders Peirce [electronic resource]***,** Charlottesville, VA, InteLex Corp., 1992

Rema Rossini Favretti, *La linguistica applicata. Aspetti. Problemi. Percorsi,* Bologna, Patron, 1998

Francesco Sabatini, Gian Tommaso Scarascia Mugnozza, *Innovazioni lessicali nell'italiano d'oggi. Riflessioni tra le "Raccomandazioni di Mannheim-Firenze" (2001) e il convegno su "lingua italiana e scienze" (2003),* in Giovanni Adamo e Valeria Della Valle e cura di, *Innovazione lessicale e terminologie specialistiche,* Firenze, Olschki, 2003, pp. 1-5

Francesco Sabatini, Vittorio Coletti, *Dizionario della lingua italiana*, Milano, Rizzoli-Larousse 2003

Giuseppe Savoca, *Lessicografia letteraria e metodo concordanziale*, Firenze, Olschki 2000

John Sinclair, *Corpus, Concordance, Collocation,* Oxford, Oxford U. P. 1991

John Sinclair, *Current Issues in Corpus Linguistics,* in *Linguistica e informatica. Corpora, multimedialità e percorsi di apprendimento,* a cura di Rema Rossini Favretti, Roma, Bulzoni, 2000, pp. 29-38

Stefania Spina, *Fare i conti con le parole. Introduzione alla linguistica dei corpora*, Perugia, Guerra, 2001

Ian Svartvik, *Corpus linguistics comes of age,* in *Directions in Corpus Linguistics,* a cura di Ian Svartvik, Berlin-New York, Mouton De Gruyter, 1992

Nicola Zingarelli, *Lo Zingarelli 2004 Vocabolario della lingua italiana*, Bologna, Zanichelli 2003